Yogesh Kodgire

Data Pipeline & Modeling Guide: Sales and Shipping Data

Contents

[**1. Overview** 2](#_Toc191276430)

[**2. Data Sources** 2](#_Toc191276431)

[**3. Data Model Tables and Specifications** 2](#_Toc191276432)

[**3.1. Customer Dimension (dim\_customer)** 2](#_Toc191276433)

[**3.2. Order Fact Table (fact\_order)** 3](#_Toc191276434)

[**3.3. Shipping Fact Table (fact\_shipping)** 3](#_Toc191276435)

[**4. ETL Pipeline Steps** 3](#_Toc191276436)

[**4.1 Extract** 3](#_Toc191276437)

[**4.2 Transform** 4](#_Toc191276438)

[**4.3 Load** 4](#_Toc191276439)

[**5. Performance Considerations** 4](#_Toc191276440)

[**6. Sample Queries** 5](#_Toc191276441)

[**6.1. Retrieve Active Customers from the USA** 5](#_Toc191276442)

[**6.2 Total Amount Spent by Each Customer** 5](#_Toc191276443)

[**6.3 Orders Pending for Delivery** 5](#_Toc191276444)

[**7. Final Thoughts** 5](#_Toc191276445)

## **1. Overview**

This document outlines the **data model design and ETL pipeline** specifications for the **Customer, Order, and Shipping datasets**. The goal is to create a **clean, structured, and efficient** data model for reporting, analytics, and business intelligence.

## **2****. Data Sources**

| **Dataset** | **Source System** | **Purpose** |
| --- | --- | --- |
| customer\_raw | Excel / CRM System | Stores customer profile details |
| order\_data | Excel / E-commerce System | Tracks customer purchases |
| shipping\_info | Excel / Shipping Database | Manages order shipment details |

## **3****. Data Model Tables and Specifications**

### **3.1. Customer Dimension (dim\_customer)**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| customer\_id | INT (PK) | Unique identifier for each customer (surrogate key) |
| source\_id | VARCHAR(50) | Original customer ID from the source system |
| full\_name | VARCHAR(255) | Cleaned and standardized full name |
| email | VARCHAR(255) | Validated email address |
| phone\_number | VARCHAR(20) | Standardized phone number format |
| age | INT | Customer age (filled with median if missing) |
| country | VARCHAR(50) | Standardized country name |
| is\_active | BOOLEAN | Indicates if the customer is currently active |
| created\_at | TIMESTAMP | Record creation timestamp |
| updated\_at | TIMESTAMP | Last update timestamp |
| valid\_from | DATE | Start date for tracking changes |
| valid\_to | DATE | End date for tracking changes (NULL for active customers) |

### **3.2. Order Fact Table (fact\_order)**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| order\_id | INT (PK) | Unique order identifier |
| customer\_id | INT (FK) | Links to dim\_customer.customer\_id |
| item | VARCHAR(255) | Product purchased |
| amount | DECIMAL(10,2) | Order amount |
| order\_date | DATE | Date of purchase |

### **3.3. Shipping Fact Table (fact\_shipping)**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| shipping\_id | INT (PK) | Unique shipping identifier |
| order\_id | INT (FK) | Links to fact\_order.order\_id |
| customer\_id | INT (FK) | Links to dim\_customer.customer\_id |
| status | VARCHAR(50) | Shipping status (Pending, Delivered) |
| shipped\_date | DATE | Date when the order was shipped |

## **4.** **ETL Pipeline Steps**

### **4.1 Extract**

* **Customer Data:** Pull records from CRM (customer\_raw).
* **Order Data:** Extract transaction details from the e-commerce system (order\_data).
* **Shipping Data:** Retrieve order shipment details from (shipping\_info).

### **4.2 Transform**

* **Data Cleaning & Standardization:**
  + Remove **duplicates**.
  + Standardize **text fields** (title case names, format phone numbers).
  + Fill missing values in **age (median)** and **categorical fields (Unknown)**.
  + Map country names to a standardized format.
* **Data Integration:**
  + Join **Customer**, **Order**, and **Shipping** datasets.
  + Assign **surrogate keys** (customer\_id, order\_id, shipping\_id).
* **Data Validation:**
  + Ensure order\_id exists before linking to shipping\_id.
  + Verify that customer records exist in the dim\_customer table.

### **4.3 Load**

* Insert transformed data into **dim\_customer**, **fact\_order**, and **fact\_shipping**.
* Apply **Slowly Changing Dimension Type 2 (SCD-2)** for tracking customer changes.
* Ensure indexes on customer\_id, email, and phone\_number for optimized lookups.

## **5.** **Performance Considerations**

* **Indexes on:** customer\_id, order\_id, email, phone\_number.
* **Partitioning on:** country in dim\_customer and order\_date in fact\_order.
* **Incremental Data Loads** to avoid full refreshes.

## **6****. Sample Queries**

### **6.1. Retrieve Active Customers from the USA**

SELECT customer\_id, full\_name, email, phone\_number

FROM dim\_customer

WHERE country = 'USA' AND is\_active = TRUE;

### **6.2 Total Amount Spent by Each Customer**

SELECT c.customer\_id, c.full\_name, SUM(o.amount) AS total\_spent

FROM fact\_order o

JOIN dim\_customer c ON o.customer\_id = c.customer\_id

GROUP BY c.customer\_id, c.full\_name

ORDER BY total\_spent DESC;

### **6.3 Orders Pending for Delivery**

SELECT o.order\_id, c.full\_name, s.status, o.amount

FROM fact\_order o

JOIN fact\_shipping s ON o.order\_id = s.order\_id

JOIN dim\_customer c ON o.customer\_id = c.customer\_id

WHERE s.status = 'Pending';

## **7****. Final Thoughts**

With this **optimized data model**, the company can ensure:

✅ **Clean and structured data for analysis**  
✅ **Faster query performance** for dashboards and reports  
✅ **Seamless integration of sales, customer, and shipping insights**